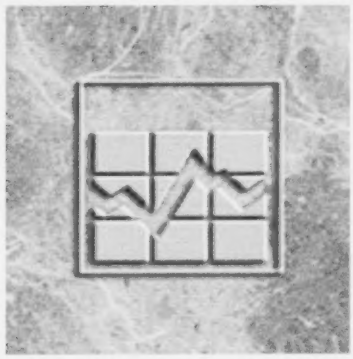
Catalogue no. 89-648-X — No. 002
ISSN 1927-0100
ISBN 978-1-100-20704-9

Research Paper

Longitudinal and International Study of Adults Research Paper Series

Historical data linkage of tax records on labour and income: The case of the Living in Canada Survey pilot



by Andrew Heisz, Manon Langevin and Jeffrey Randle

Income Statistics Division, Statistics Canada,
5th Floor, Jean Talon Building, 170 Tunney's Pasture Driveway,
Ottawa, Canada, K1A 0T6

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at infostats@statcan.gc.ca,

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-877-287-4369

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

To access this product

This product, Catalogue no. 89-648-X, is available free in electronic format. To obtain a single issue, visit our website, www.statcan.gc.ca, and browse by "Key resource" > "Publications."

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "About us" > "The agency" > "Providing services to Canadians."

Published by authority of the Minister responsible for Statistics Canada

© Minister of Industry, 2013

All rights reserved. Use of this publication is governed by the Statistics Canada Open Licence Agreement (http://www.statcan.gc.ca/reference/licence-eng.html).

Cette publication est aussi disponible en français.

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^s	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Abstract

Data matching is a common practice used to reduce the response burden of respondents and to improve the quality of the information collected from respondents when the linkage method does not introduce bias. However, historical linkage, which consists in linking external records from previous years to the year of the initial wave of a survey, is relatively rare and, until now, had not been used at Statistics Canada. The present paper describes the method used to link the records from the Living in Canada Survey pilot to historical tax data on income and labour (T1 and T4 files). It presents the evolution of the linkage rate going back over time and compares earnings data collected from personal income tax returns with those collected from employers' file. To illustrate the new possibilities of analysis offered by this type of linkage, the study concludes with an earnings profile by age and sex for different cohorts based on year of birth.

Table of content

1. Introduction 7
2. Linkage 7
3. Data source 8
4. Analytical results 8
4.1 Linkage rate between 1990 and 2007 8
4.2 Comparison of earnings from T1 and T4 files 10
4.3 Profiles of earnings by age and sex 11
5. Conclusion 12
Bibliography 13

List of tables and figures

Table 2.1-1 Permission for yearly and historical linkage from the LCS pilot questionnaire 8
Figure 4.1-1 Linkage rate of the personal income tax return file (T1) from 1990 to 2007 9
Table 4.2-1 Difference in earnings reported in the T1 file and the T4 file 10
Table 4.2-2 Median employment earnings according to statements in the T1 and T4 files 10
Figure 4.3-1 Earnings profile by age and sex (Men) 11
Figure 4.3-2 Earnings profile by age and sex (Women) 12

1. Introduction

Data linkage is a common practice at Statistics Canada and in several other statistical agencies around the world. It is a good way to reduce the costs associated with survey activities and to enhance the analytical strength from existing data sources. Certain types of information will always be hard to collect by survey, either because it requires significant recall by respondents (e.g. monthly calendar of employment status) or because the nature of the subject may be embarrassing to discuss with a stranger (e.g. victim of sexual assault). For this reason, analysts in several fields are regularly confronted with insufficient data, which limits the type of analysis that can be conducted.

Data linkage can also be used, in the context of social surveys, to replace questions related to income, significantly reducing the response burden of respondents who consent to this type of procedure and increasing the accuracy of the data collected. The Survey of Labour and Income Dynamics is a good example of this. Generally, 80% of respondents of this survey consent each year to data matching. The consent is valid for the duration of the respondents' participation in the survey and the linkage of records is carried out solely between data files belonging to the same collection year. However, the linkage carried out with the pilot of the Living in Canada Survey (LCS) is somewhat different because it involves historical linkage. More specifically, the LCS records were linked to tax files from years prior to the LCS collection year; this type of linkage had never been done before with a Statistics Canada social survey.

The purpose of this study is to explain how the historical linkage was done and to present the benefits in terms of analytical potential. In particular, the study examines (1) the degree to which the linkage rate diminishes going back in time, (2) the accuracy of the information contained in the different linked tax files, and (3) the potential of the retrospectively linked information in analyzing phenomena that require a long data series.

2. Linkage

The Living in Canada Survey (LCS) pilot data were linked with different tax files of individuals and businesses: (i) the personal income tax form (T1 file), (ii) the statement and summary of compensation paid by employers (T4 file), and (iii) the Pension Plans in Canada Survey file[1]. For these files, two different types of linkage were carried out: (i) a yearly linkage (renewable for each new wave of the survey) and (ii) an historical linkage of tax data for the years going back to 1990.

Despite the fact that this second type of matching had never been done before at Statistics Canada, there is some emerging literature on the topic (Reimer and Künster, 2004; Roemer, M., 2002; Sears and Rupp, 2003). A team of German researchers had carried out an exercise similar to ours by comparing administrative data on employment status to employment data from a survey (Huber and Schmucker, 2009). One of their main findings was that the probability of the two files corresponding was negatively correlated to the number of events. In other words, they found that respondents who experienced several changes in employment status during the period analyzed were significantly more likely to forget events and to produce an incorrect statement (memory bias). This appears to indicate that the historical administrative data improved the quality of the information for this type of individual in particular.

The consent of LCS respondents to yearly and historical linkage was requested completely separately. The first question allowed respondents to give permission to access their tax file for the pilot's reference year and for the duration of their participation in the survey. Permission to access the same files, but for all calendar years going back to 1990, was asked in a second question and solely of those respondents who had previously consented to linking their tax files for 2007 (Table 2.1-1).

[1] The Pension Plans in Canada Survey is a complete annual survey of registered pension plans in Canada. It includes information on the various terms and conditions of these plans, membership and contributions.

Table 2.1-1 Permission for yearly and historical linkage from the LCS pilot questionnaire

In order to increase the accuracy of the data and reduce the time of the interview, you can give permission for Statistics Canada to access your income tax records.

1. Statistics Canada would obtain information from: the T1 Income Tax and Benefit Return, the T4 file of records from employers, and the employer pension plan file. The information we obtain would be used for statistical purposes only, and would be kept confidential. Do we have this permission?

2. Statistics Canada would also like your permission to access information contained in your past tax records back to 1990. This will enhance the quality of the results from the survey. The information we obtain would be used for statistical purposes only, and would be kept confidential. Do we have this permission?

Source: Questionnaire of the Living in Canada Survey pilot (2007)

For confidentiality reasons, respondents could not be asked directly for their social insurance number (SIN) in the LCS interview. Consequently, respondents' SIN was identified by probabilistic record linkage and by using auxiliary variables such as last name, first name, sex, date of birth, marital status, address and postal code. It should be noted that for technical reasons, data from T4 files were not available for years prior to 2001 at the time of this analysis.

3. Data source

The main purpose of the Living in Canada Survey pilot (LCS) was to provide Canada with a longitudinal survey with panels of indefinite length and content covering several topics of interest for the development of public policy. More specifically, the survey's content should enable researchers to identify the links between the four main areas of an individual's life, namely, work, education, family life and health status. Similar surveys are available in several countries and have proven useful in developing better policies and programs by all levels of government.

The pilot was a means to test the survey questionnaire. Sample coverage was limited to four of Canada's ten provinces: New Brunswick, Quebec, Ontario and Saskatchewan. The pilot covered the population of these four provinces, excluding regular members of the Canadian Forces, individuals living in institutions and on a native reserve. In total, the file contains data from 2,881 respondents, aged 15 years and older, 79% of whom consented to linking their data to the reference year. Of that number, 94% also consented to the historical linkage of their tax data going back to 1990. Since the analysis focused only on respondents who consented to both types of linkage, the sample available included 2,137 observations.

4. Analytical results

4.1 Linkage rate between 1990 and 2007

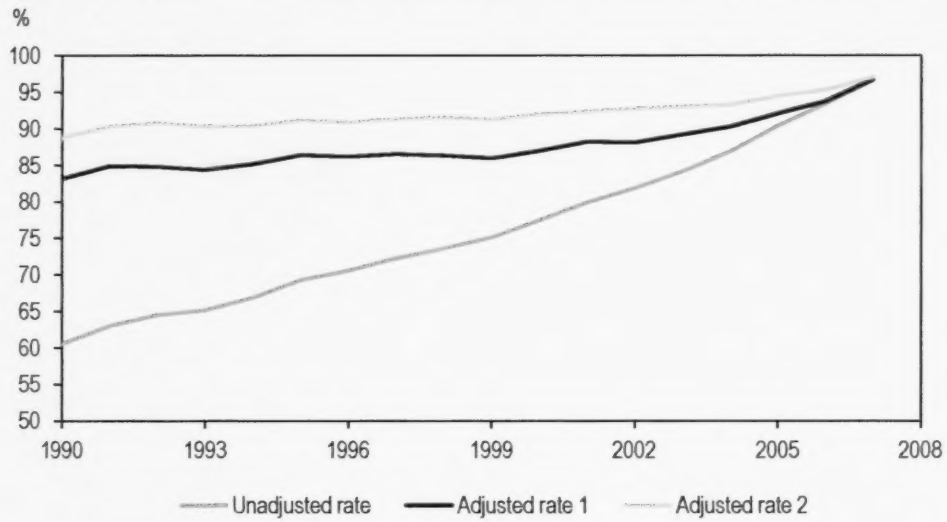
Data linkage fails when the unique matching key does not find a match in the linked file. There may be several reasons for this failure, and in the case of historical linkage, the potential sources are even more numerous. Other than errors arising from the processing of records, the extent of which is difficult to quantify, some of the reasons for a failed link are more easily identifiable and can be divided into two main sources: (1) the person did not file an income tax return or did not work for an employer during the taxation year; or (2) the person's matching key was not constant over time.

This second type of reason will introduce bias in the longitudinal analysis of data since the linkage failure cannot be attributed to the absence of an income tax return during those years. Although the Social Insurance Number (SIN) is a relatively stable key over time, it can change, notably in the case of immigrants who are assigned a temporary SIN on their arrival in Canada and who are then assigned a permanent SIN[2].

The linkage rate for the T1 data files was calculated for the period from 1990 to 2007 to determine the degree to which it declines going back over time. Three different types of rates were calculated: (1) a gross rate using all of the available sample, (2) an adjusted rate using a sample excluding respondents aged 15 years or under and immigrants landed in, or in the year prior to, the fiscal year, and (3) a second adjusted rate based on a sample excluding respondents aged 20 years or under and immigrants landed in the three years preceding the fiscal year. The restrictions on age and immigrant status reflect the fact that these two groups are less likely to produce an income tax return during a given year or to have a constant SIN over time.

As Figure 4.1-1 shows, the results indicate that the linkage rate drops going back in time regardless of the sample used for the calculation. However, the most significant reduction in the linkage rate occurs when the calculation is based on a sample without exclusions, in which case it drops from 96.7% in 2007 to only 60.5% in 1990. When we exclude those respondents for whom linkage was *a priori* more likely to fail, the rate drops much less quickly and remains at over 80% in 1990 for both samples with exclusions.

Figure 4.1-1 Linkage rate of the personal income tax return file (T1) from 1990 to 2007



Source: LCS (2007) and linked data from T1 file (1990 to 2007)

[2] It is difficult to assess the longitudinal bias introduced by retrospective linkage of immigrants because there are several possible reasons for the failure of their linkage. Because they are likely to be assisted by a family member upon their arrival in Canada or to leave the country several times after their immigration, it is difficult to know if their data are missing because they did not file an income tax return during these years (no introduction of bias) or because of a change in SIN over time.

4.2 Comparison of earnings from T1 and T4 files

One way to verify the accuracy of the information reported by a respondent on a survey is to compare it to information contained in administrative files. Administrative files are considered a more reliable data source and many studies use this type of methodology to assess the impact of non-response and attrition on estimate quality (Pyy-Martikainen and Rendtel, 2003; Roemer, 2002; Johnson and McMahon, 2002). However, very few studies make this type of comparison among different administrative files in order to determine the correspondence between the amounts reported in each source.

Earnings amounts in T1 data files and in T4 data files were compared for the period from 2001[3] to 2007. The results show that the majority of cases—approximately 98% each year—present a similar earnings situation in both the personal file (T1) and the employer file (T4) [Table 4.2-1]. In other words, when earnings are found in the T1 file, they are also found in the T4 file and vice-versa. In 2007, for example, there are only 33 cases with earnings from a single data source. Of that number, the majority of linked information, some 21 cases, comes from the T1 file and the median earnings associated with those cases is only one dollar (Table 4.2-2).

Table 4.2-1 Difference in earnings reported in the T1 file and the T4 file

	2007	2001
	%	
T1 earnings between (T4 earnings – $1000.00 and T4 earnings + $1000.00)	98.66	98.83
T1 earnings between (T4 earnings – $1.00 and T4 earnings + $1.00)	95.17	87.59
T1 earnings less than (T4 earnings – $1000.00)	0.79	0.90
T1 earnings greater than (T4 earnings + $1000.00)	0.55	0.27

Source: LCS (2007) and linked data from T1 and T4 files (2001-2007)

When employment earnings are present in both files, the difference between the amounts reported is no more than one dollar in 95% of cases (Table 4.2-1). The reason for this finding is quite simple and is due to the fact that earnings amounts reported in the T1 files are rounded to the nearest dollar, which is not done with the amounts reported in the T4 files. Referring to the results in Table 4.2-2, we find that the median earnings are very similar from one data source to the other. From 2001 to 2007, the difference in median employment earnings calculated from the two data sources is on average $150.00.

Table 4.2-2 Median employment earnings[4] according to statements in the T1 and T4 files

	2007		2001	
	dollars		dollars	
	T1	T4	T1	T4
Statement of earnings in T1 and T4 files	31,745	31,886	30,774	30,930
Statement of earnings in T1 file only	1.00	n/a	1.14	n/a
Statement of earnings in T4 file only	n/a	278	n/a	462

Source: LCS (2007) and linked data from T1 file and T4 file (2001-2007)

[3] At the time of this analysis, information from T4 files was not available for years prior to 2001.

[4] Employment earnings expressed in 2007 constant dollars.

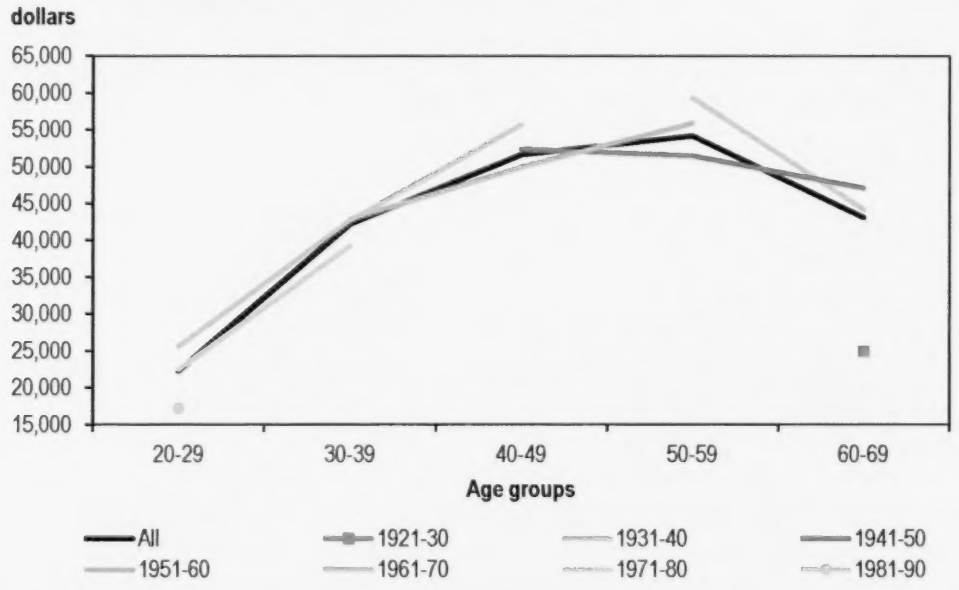
4.3 Profiles of earnings by age and sex

The lack of longitudinal data on certain types of research topics leads researchers to use more approximate analytical methods to examine certain phenomena. For example, creating synthetic cohorts is a common practice to monitor the evolution in earnings over time when data comes from cross-sectional surveys. Retrospectively linked data could overcome this data limitation and also allow analysis of any issue requiring the use of a long data series.
To fully exploit the potential of linked information and to verify whether such information can produce analytical results comparable to those achieved using other data sources, an earnings profile by age and sex was developed for different cohorts based on birth year. Given that the earnings amounts reported were similar regardless of the source, the earnings reported in personal tax return files were used to develop the earnings profile, thus producing a longer data series than the data drawn from the employers' file. The sample was divided into seven birth year groups, each with a 10-year interval, for which the change in employment earnings for different age groups was tracked.

As Figures 4.3-1 and 4.3-2 show, the earnings profile developed using historical data reveals general trends comparable to those from other data sources, specifically: (i) an earnings profile increasing with age and beginning to decline around age 50 years as individuals retire from the workforce, and (ii) a higher earnings profile for men than for women for all age groups analyzed.

The earnings profile also shows trends by cohort similar to those found in related literature (Beach and Finnie, 2004; Boudarbat, Lemieux and Riddell, 2003; Burbidge, Magee and Robb, 1997). First, there is a trend of lower earnings at the beginning of their career for workers in the most recent cohorts (e.g. 1961-70 cohort versus 1971-80 cohort) and second, a faster growth in earnings[5] for workers in the more recent cohorts. As a result, there is a trend toward higher earnings at career end for workers in the more recent cohorts than for the older cohorts (e.g. 1951-60 cohort versus 1941-50 cohort).

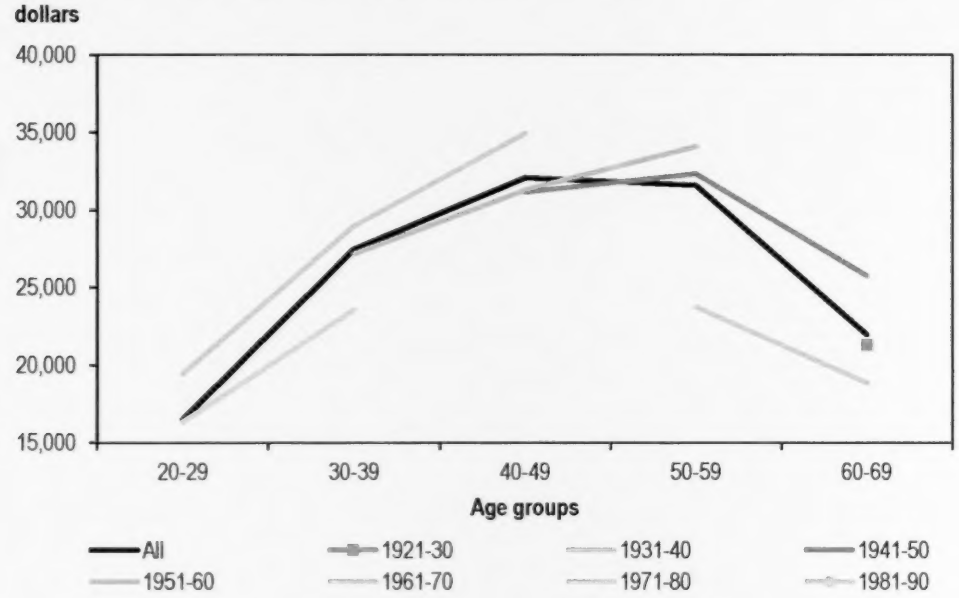
Figure 4.3–1 Earnings profile by age and sex (Men)



Source: LCS (2007) and linked data from T1 files from 1990 to 2007

[5] In the literature, this growth is often associated with the growth in returns to education.

Figure 4.3–2 Earnings profile by age and sex (Women)



Source: LCS (2007) and linked data from T1 files from 1990 to 2007

5. Conclusion

The results of this study represent only the first step in our research. At present, the preliminary findings indicate that the information provided by retrospectively linked data could be of sufficiently good quality to be used in analyzing phenomena that require the availability of a long data series. According to several studies, this type of linkage could also gather more reliable data for respondents likely to provide incorrect responses on their historical information (e.g. workers with several jobs in recent years).

In the specific case of the Living in Canada Survey, the historical linkage of tax data also made it possible to add complementary information to the survey data about the respondents' past (education, family life, work and health). The addition of this historical information allows analysis of longitudinal phenomena as of the first wave of the survey, an operation that would not normally be possible until after several waves. It also offers new tools to analysts without increasing the response burden of respondents or the costs associated with the survey's activities.

Bibliography

Beach, C. and Finnie, R. (2004), "A Longitudinal Analysis of Earnings Change in Canada", Analytical Studies BranchResearch Paper, Statistics Canada.
http://www.statcan.gc.ca/pub/11f0019m/11f0019m2004227-eng.pdf

Burbidge, J.B., Magee, L. and Robb, A.L. (1997), "Cohort, year and age effects in Canadian wage data", Independence and Economic Security of the Older Population (IESOP), McMaster University, Research Paper no.19. http://www.ciln.mcmaster.ca/papers/cilnwp13.pdf

Boudarbat, B., Lemieux, T. and Riddell, W.C. (2003), "Recent Trends in Wage Inequality and the Wage Structure in Canada", University of British Columbia, TARGET, Research Paper no.6.
http://faculty.arts.ubc.ca/criddell/rectrend.pdf

Huber, M. and Schmucker, A. (2009), "Identifying and Explaining Inconsistencies in Linked Administrative and Survey Data: The Case of German Employment Biographies", Historical Social Research, Vol. 34 - 2009 - No. 3, 230-241.

Johnson, B.W. and McMahon, P.B. (2002), "Using auxiliary information to adjust for non-response in weighting a linked sample of administrative records", Internal Revenue Service, Presentation to the 2002 «*American Statistical Association*».
http://www.amstat.org/sections/srms/Proceedings/y2002/Files/JSM2002-001045.pdf

Pyy-Martikainen, M. and Rendtel, U. (2003), "The Effects of Panel Attrition on the Analysis of Unemployment Spells", CHINTEX, Research Paper no.10.
http://www.destat.net/jetspeed/portal/cms/Sites/destatis/Internet/DE/Content/Wissenschaftsforum/Chintex/Projekt/Downloads/WorkingPaper6_042003,property=file.pdf

Reimer, M. and Künster, R. (2004), "Linking Job Episodes from Retrospective Surveys and Social Security Data: Specific Challenges, Feasibility and Quality of Outcome", Berlin, Max-Planck-Institut für Bildungsforschung.
http://www.mpib-berlin.mpg.de/en/institut/dok/full/reimer/linkingj/Arbeitsber8.pdf

Roemer, M. (2002), "Using Administrative Earnings Records to Assess Wage Data Quality in the March Current Population Survey and the Survey of Income and Program Participation", U.S. Census Bureau, LEHD Program, Technical Paper no. TP-2002-22.
http://lehd.did.census.gov/led/library/techpapers/tp-2002-22.pdf

Sears, J. et Rupp, K. (2003), "Exploring Social Security Payment History Matched with the Survey of Income and Program Participation", presented on the 18th of November at the «*Federal Committee on Statistical Methodology*» de 2003.
http://www.oecd.org/dataoecd/15/36/36232603.pdf